Automated Interpretation of Lung Ultrasound for COVID-19 and Tuberculosis diagnosis

Chloé Soormally

Master's thesis
2023:E45

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Centrum scientiarum mathematicarum

MASTER'S THESIS REPORT

Automated Interpretation of Lung Ultrasound for COVID-19 and Tuberculosis diagnosis

Chloé Soormally

Supervision

Mikael Nilsson (LTH)
Mary-Anne Hartley (EPFL)

LUND UNIVERSITY

LTH

FACULTY OF ENGINEERING

Abstract

BACKGROUND. Early and accurate detection of infectious respiratory diseases like COVID-19 and tuberculosis (TB) plays a crucial role in effective management and the reduction of preventable mortality. However, molecular diagnostic tests for these infections are expensive and not easily implementable in resource-limited settings, which suffer the majority of the burden. Lung Ultrasound (LUS) presents a cost-effective alternative for disease detection at the point of care, and its potential can be enhanced through automation using deep learning techniques to overcome the challenges of difficult image interpretation. *DeepChest*, a neural attention network, has been designed to predict the diagnosis of COVID-19 from LUS images and has shown promising results.

AIM. This study aims to further explore the predictive capabilities of *DeepChest* with an out-of-distribution dataset and extend its application to TB diagnosis.

METHODS/FINDINGS. For COVID-19 (resp. TB), this study is based on a main dataset and an out-of-distribution dataset consisting of patients attending an emergency department in Switzerland (resp. an outpatient facility in a TB-endemic region) between February 2020 and March 2021 (resp. between October 2021 and May 2023) with suspected COVID-19 (resp. TB) pneumonia and ground truth labels are RT-PCR.
To assess the generalizability of *DeepChest* for COVID-19 diagnosis, the model trained on the main dataset (296 patients, still LUS images) was tested on an out-of-distribution dataset (135 patients more severely affected, mainly frames extracted from LUS videos). We found that the performance on the out-of-distribution dataset was poor. However, by fine-tuning *DeepChest* on the latter, the best performance was achieved when using three random frames per video (AUC ROC 0.84 +/- 0.03) instead of one (AUC ROC 0.78 +/- 0.06).
To assess the performance of *DeepChest* for TB diagnosis, the model was trained on the main dataset (386 patients, still LUS images collected in an urban area of Benin). Its performance (AUC ROC 0.92 +/- 0.02) outperformed the LUS expert baseline (AUC ROC 0.84 +/- 0.01) and the clinical baseline (AUC ROC 0.89 +/- 0.01). Additionally, a multimodal model incorporating clinical data alongside LUS images was developed. It achieved the best classification performance (AUC ROC 0.94 +/- 0.01). However, when tested on an out-of-distribution dataset (150 patients, still LUS images collected in a rural region of South Africa, with much more severe presentation), the generalizability of *DeepChest* was found to be low (AUC ROC 0.64 +/- 0.03).

CONCLUSION. The findings of this study are promising, demonstrating the potential of *DeepChest* for COVID-19 and TB diagnosis using LUS images. Poor generalization to populations with more severe forms of the diseases shows the importance of either collecting more representative samples or ensuring that implementation is constrained to the target population.

Acknowledgment

I would like to express my sincere gratitude to Annie Hartley for her invaluable support and guidance during my time at EPFL. Being hosted in her Intelligent Global Health (iGH) research group has been an incredible opportunity, allowing me to contribute to an exceptionally interesting and challenging project.

I would like to extend my thanks to Mikael Nilsson whose contributions and insightful suggestions have greatly complemented Annie's guidance, enriching the project with diverse perspectives and approaches on numerous occasions.

Last but not least, I am grateful to all the individuals without whom this project would not have been possible. My deepest thanks go to the teams involved in data collection in Switzerland, Benin and South Africa and, of course, to all the patients who participated in the studies. Their involvement and efforts have been instrumental in advancing our understanding and improving global health outcomes.

Contents

1. Introduction

Background

Tuberculosis and COVID-19 - the importance of early, rapid and accurate detection

The rapid and unexpected emergence of COVID-19 in 2020 highlighted the lack of preparedness of healthcare systems globally. Early and rapid detection of positive cases of a disease with epidemic potential is crucial for effective control and management, as it allows to identify, isolate and treat appropriately, which can in turn help limit the spread of the virus and reduce the mortality rate [1]. During the pandemic, the diagnosis of COVID-19 relied essentially on two types of tests: molecular tests (such as RT-PCR) and antigenic tests. If the first type has become the gold standard for the detection of the infection due to its high specificity, its use is limited by its higher cost, need for specific infrastructure and expertise, as well as by the limited availability of necessary reagents [2]. Antigenic tests, on the other hand, offer a cheaper alternative to molecular testing and are therefore more accessible in resource-limited settings but have the significant disadvantage of being less accurate [3] as well as relying on a supply chain and creating significant plastic pollution.

The challenges of disease control are not unique to COVID-19 and are in many ways reminiscent of those faced in the current fight against another global pandemic, that of tuberculosis (TB), an infectious disease caused by airborne mycobacteria that primarily affects the lungs. Despite being preventable and curable, TB still causes 1.6 million deaths each year worldwide, with low and middle-income countries being the most affected [4]. Increasing drug resistance has made TB a re-emerging global health threat. The World Health Organization (WHO) puts early detection and treatment of TB at the center of its strategy to eradicate TB in the coming years [5]. WHO's current recommendations for TB detection tools include the automated molecular tests - Xpert MTB/RIF and Xpert Ultra [6] - and chest radiography (CXR) [7]. However, the use of these tests in the places where they are most needed - namely TB-endemic regions - is limited by a number of factors such as the availability of laboratory facilities and expertise, delays and traceability issues due to sample transport or consumable stock-outs. CXR, on the other hand, requires expensive equipment, trained personnel and exposes patients to radiation.

For management of diseases with epidemic potential, there is a critical need to develop diagnostic tools that are accessible for low-resources or remote settings, that do not rely on costly infrastructure, specific expertise or limited consumables, and that are ideally implementable at the point-of-care.

Lung Ultrasound (LUS) - a promising approach for TB and COVID-19 detection

Lung Ultrasound (LUS) is a non-invasive and virtually consumable-free imaging technique, which is safer and more affordable than CXR. Several recent studies have shown that LUS is a powerful tool for detecting various respiratory syndromes and lung abnormalities such as pneumonia, pleural effusion or pulmonary edema [8, 9, 10].

Recently, ultrasound systems have been developed into portable and cost-effective pocket-sized devices, pluggable into a mobile phones. This transformation has made LUS particularly interesting in remote or low-resource areas [11] and it has been proposed as a potential tool for the management of tuberculosis in endemic areas [12]. Its popularity grew during the COVID pandemic, where several studies showed its promise for COVID-19 diagnosis at the point-of-care [13].

However, while it has been established that LUS can reveal certain lung abnormalities, diagnosis remains imperfect as image interpretation can be difficult due to strong operator dependency and lack of standardization of acquisition protocols. Moreover, it has been shown that in developing countries, the lack of training of sonographers was one of the major obstacles to the use of ultrasound, more than the cost of machines and maintenance [14].

Automating LUS with Deep Learning

Deep learning is a possible way to address these challenges. The automation of ultrasound using deep learning could improve diagnosis by providing a more objective interpretation of the images and overcoming human limitations in recognizing disease-specific patterns. In addition, the use of deep learning could democratize the practice of ultrasound to non-specialist health workers, which is particularly useful in cases of limited medical resources, during pandemics or in remote and low-resource areas.

The emergence of a multitude of studies aiming to automate the analysis of COVID-19 LUS images with deep learning techniques for classification or segmentation has been driven by the 2020 pandemic [15]. However, despite the considerable volume of research, numerous challenges remain. Concerns regarding the generalizability of the proposed methods arise due to limitations observed in these studies. These limitations encompass issues such as the size and quality of datasets, which often lack standardized acquisition protocols to minimize biases, the quality of manual labeling, which can be operator-dependent, the relevance of classification tasks that may not reflect real-world scenarios (e.g., COVID-19 positive vs. healthy discrimination), and the use of algorithms that operate at the image level rather than the patient level, thereby reducing the clinical applicability of the methods. In contrast to the extensive focus on deep learning applied to LUS images for COVID-19 detection, the automation of LUS interpretation for TB detection has not received similar attention despite the ongoing pandemic.

In the context of the COVID-19 pandemic, a deep learning model - *DeepChest* [16] - was developed in the iGH laboratory to automate the detection of COVID-19 from LUS images. Unlike most models from the literature, *DeepChest* aggregates the input images to make a prediction at the patient level. To ensure the quality of the study, the data was collected from a strict and standardized protocol on a cohort of nearly 300 patients presenting to a Swiss emergency department with suspected COVID-19 pneumonia between February 2020 and March 2021. *DeepChest* was shown to achieve a balanced accuracy of nearly 80% on the test set, well above the clinical and LUS expert models which do not reach 70% of balanced accuracy.

Aim and objectives

The promising results of *DeepChest* motivate the aim of this thesis to investigate further its predictive potential for COVID-19 diagnosis with an out-of-distribution dataset and to extend its application to TB diagnosis, another respiratory disease that would also benefit from a rapid and accurate diagnostic tool such as *DeepChest*.

Objectives for COVID-19 are:

1. To understand and replicate the results of *DeepChest* on the original COVID-19 dataset consisting of patients presenting to an ER in Switzerland during the COVID-19 pandemic.

2. To explore ways to preprocess the images from an out-of-distribution COVID-19 dataset that contains images collected from more severely ill patients that were hospitalized during the COVID-19 pandemic.
3. To explore the generalizability and performance of *DeepChest* on this dataset.

Objectives for TB are:

1. To preprocess the datasets relevant to TB diagnosis and create *DeepChesTB*.
2. To compare the performance of *DeepChesTB* for the diagnosis of TB with a clinical model and a LUS expert human model
3. To build a multimodal model using both clinical data and LUS images as input and compare the performance with previous models.
4. To assess the performance of *DeepChesTB* on an out-of-distribution dataset of more severely ill patients.

Related work

Deep learning (DL) in medical imaging is a very broad research topic. The imaging modalities used for medical diagnosis are numerous and include among others ultrasound, radiography, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). The use of DL in medical image analysis are also varied, including image/lesion classification, organ/lesion detection and organ segmentation. The areas of anatomical application are also increasingly diverse, and recently included the lung [17]. We narrow our focus on studies around lung ultrasound or chest radiography, excluding modalities that are too expensive and/or unsuitable for the diagnosis of respiratory pathologies in low-resource settings.

Deep learning and Chest radiography

According to the review by Erdi Çallı et al. [18], the literature around deep learning applied to CXR is dominated by studies on performed on ChestX-Ray14 dataset (aiming to classify 14 lung pathologies), where the detection of pneumonia, pneumothorax, TB or COVID-19 are among the most covered topics.

In their review, Mustapha Oloko-Oba et al. [19] present state-of-the-art deep learning methods for TB detection from CXR. They note that most of the studies propose models based on existing and often pretrained CNN architectures such as ResNet, VGG, Inception, or AlexNet. Moreover, it appears from the review that a large majority of these studies are based on only two public datasets, namely Montgomery and Shenzen [20], which contain respectively 138 CXR images (80 normal / 58 manifestation of TB) and 662 CXR images (323 normal / 336 manifestation of TB). The performances achieved on these datasets are often very impressive, with accuracy reaching 99%, however they should be interpreted with caution. While the aforementioned models demonstrate excellent discrimination between TB and healthy cases, their performance in differentiating TB from other lung infections in a real-world scenario remains uncertain. Additionally, the label of "TB manifestation" is a human label, which in itself has limited performance.

The issue of dataset standardization is crucial when interpreting the reported results. For example, Tawsifur Rahman et al. [21] released a large dataset of 7000 CXR images (3500 normal/3500 TB), constructed from 4 public datasets. However, a major bias exists as the TB images and the normal

images are sourced from different datasets. It is therefore difficult to know whether the trained models learn to recognize the differences in patterns between healthy vs TB or simply the differences in image acquisition.

The results reported in the literature should therefore be considered in light of these limitations, highlighting the need for new studies based on datasets collected with protocols that minimize acquisition bias and are oriented towards more realistic applications.

Beyond the limitations in terms of datasets, it is worth recalling that CXR is not the most cost-effective imaging modality for low-resource settings and exposes patients to radiation, making it unsuitable for the most vulnerable populations and precluding regular use e.g., to monitor disease progression.

Deep Learning and Lung Ultrasound

LUS offers several advantages over CXR, including cost-effectiveness, enhanced safety, and improved accessibility. As for CXR, leveraging its potential can be achieved through automation using deep learning techniques.

In response to the pandemic in 2020, there has been a growing interest in automating the analysis of LUS images for the detection of COVID-19 [15]. Born et al. [22] were among the first to work on the automatic detection of COVID-19 from ultrasound images. They developed a deep learning model, POCOVID-Net, adapted from VGG16 and trained on a dataset of 64 videos that classifies three classes - namely COVID-19 pneumonia, non-COVID-19 pneumonia, and healthy. The model achieved 82% balanced accuracy with a sensitivity of 96% for the class COVID-19 pneumonia. Diaz-Escobar et al. [23] did a similar study on a larger version of this dataset (185 videos and 68 images) and compared the performance of several preexisting and pretrained models (VGG19, InceptionV3, Xception, and ResNet50) and showed that InceptionV3 gave the best performance with a balanced accuracy of 89.1% (and AUC ROC of 97.1%).

The two previous studies performed analysis at image-level (often frames extracted from video) but some works have focused on analysis from videos directly. Ebadi et al. [24] proposed a model based on a preexisting DL model for human motion/activity classification (Kinetics-I3D network) to detect certain features in COVID-19 videos. As Ebadi et al., other studies have sought to detect pathology features rather than simply classifying COVID-positive and COVID-negative cases. For example Subhankar Roy et al. [25] proposed an approach to assess the severity of the condition using a scoring system, which could facilitate the use of LUS for disease progression monitoring and patient triage.

However, similar to studies based on CXR images, limitations regarding dataset size and quality exist in LUS studies, raising concerns about the generalizability of the models. Furthermore, despite extensive research on deep learning applied to CXR for TB detection, there is a noticeable lack of studies focusing on automating LUS for TB detection.

The remarkable outcomes of recent studies suggest that deep learning can enhance the potential of LUS for COVID-19 detection. This paves the way for future research based on larger standardized datasets for the diagnosis of COVID-19 and other infectious pulmonary diseases, such as TB.

2. Methods

2.1 Datasets

Datasets description

For each of the two pathologies, data from two cohorts are available. The data were collected according to strict and standardized protocols designed to build clean and standardized datasets [26, 27]. The description of the different datasets is presented in Table 2.2.

Table 2.2 – Datasets description

Cohort	**COVID-19**		**TB**	
	ER	**Hospitalized**	**Benin**	**South Africa**
# of patients	296	135	386	150
Labels (% positive)	RT-PCR (70%)	RT-PCR (40%)	Xpert MTB/RIF (36%)	Xpert MTB/RIF (25%)
US imaging system	ButterflyIQ[1]	Sonosite[2] (X-porte and M-turbo)	ButterflyIQ	ButterflyIQ
# (anatomical) sites	10	10	14	14
# of images	4363	532 (X) / 107 (M)	10227	2353
# of videos	0	1692 (X) / 255 (M)	0	0
Clinical data available ?	Yes	No	Yes	Yes

[1] `https://www.butterflynetwork.com/iq-ultrasound-individuals`, [2] `https://www.sonosite.com/`

Cohorts

For COVID-19 diagnosis, data was collected from two patient cohorts. The main cohort consists of adult patients with suspected COVID-19 pneumonia attending a Swiss emergency department (ER) between February 2020 and March 2021. Patients were excluded if the LUS examination could not be performed within 24 hours of admission. The second cohort, for external validation, consists of adult patients with suspected COVID-19 pneumonia who were hospitalized in the internal medicine ward during the pandemic in 2020. The second cohort is expected to contain more severely affected patients but unfortunately, we do not have more details on the recruitment criteria for this cohort.

For TB diagnosis, data was collected from two patient cohorts consisting of adult patients attending an outpatient department between October 2021 and May 2023 with suspected pulmonary tuberculosis in a TB-endemic region. The main cohort was recruited in an urban region of Benin while the cohort for external validation was recruited in a rural region of South Africa. The latter cohort is expected to contain patients more severely ill due, in part, to a significantly higher prevalence of HIV in the population.

Ground-truth labels

The ground-truth labels are the diagnosis of the disease studied i.e., COVID+/COVID- or TB+/TB-, obtained with the RT-PCR test result for COVID-19 and the Xpert MTB/RIF test result for TB.

Anatomical sites (views)

For the COVID-19 datasets, when possible, 10 views from 10 anatomical sites were collected (anterior, posterior and lateral). For the TB dataset, when possible, 14 views from 14 anatomical sites were collected (anterior, posterior, lateral and apical). The localization in the thorax of the different anatomical sites is presented in Figure 2.1.

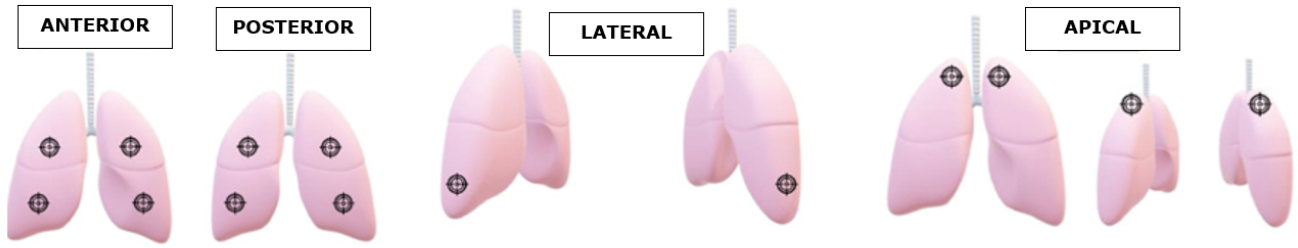


Figure 2.1 – Localization of the different possible views for the LUS images/videos.

It should be noted that sometimes it was impossible to collect an image/video of a certain anatomical site from a patient. It also happened that several images/videos of the same site were collected for the same patient. The number of images/videos per patient as well as the number of images/videos per anatomical site may thus vary.

Clinical tabular data

In addition to LUS images/videos, clinical data was also collected on the TB cohorts. A description of these data is given in Table 2.3.

Table 2.3 – Clinical data. Variables written in bold are considered categorical while the others are considered numerical (continuous).

Demographics	**sex**, age
Past medical history	**hiv** (known), **diabetes**, **lung diseases**, **previous tb**, **smoker**, **hypertension**, **cardiopathy**
Signs	temperature, bmi, heart rate, blood pressure, respiratory rate, oxygen saturation, **general state**
Lab results	**hiv**
Expert interpretation [1]	**lung ultrasound images**

[1]More details are given in the Appendix in Table A.2.

2.2 Data preprocessing

The LUS images/videos of the datasets - with the exception of the Hospitalized dataset - are collected using the ButterflyIQ imaging system. The Hospitalized dataset images/videos are collected with the Sonosite X-porte and M-turbo systems. In this sub-section, the preprocessing pipelines associated with the different imaging systems are presented, with the ButterflyIQ system on one side and the Sonosite systems on the other.

2.2.1 ButterflyIQ imaging system

Images from the ER dataset (COVID-19) were already preprocessed. We have therefore only carried out preprocessing for the images in the TB dataset. For this purpose, we followed closely the preprocessing pipeline that was designed for the ER dataset and that is presented in Figure 2.2.

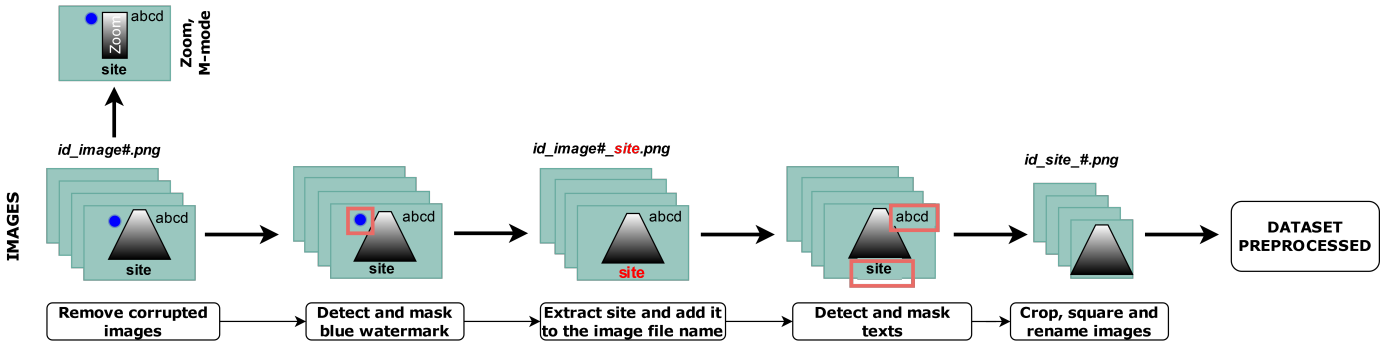


Figure 2.2 – Preprocessing pipeline for the images collected with the ButterflyIQ imaging system.

- Dataset is manually inspected and corrupted images are removed as they might introduce a bias to the dataset [1]. Examples of such corrupted images are presented in the Appendix (Figure A.1).
- The blue watermark on images is detected using template matching and masked.
- Position sites - indicating where the images were collected and written at the bottom of the latter - are extracted using Python OCR library *Pytesseract*.
- Texts on images are detected using CRAFT text detector [28] and masked.
- Images are cropped, squared, and renamed.

An overview of an image before and after preprocessing is shown in Figure 2.3.

2.2.2 Sonosite imaging systems

For the Sonosite M-turbo and Sonosite X-porte imaging systems, we designed the preprocessing pipeline as no previous work had been done on the Hospitalized dataset. An overview of this pipeline is presented in Figure 2.4.

Selecting frames: strategies

Unlike the datasets from ButterflyIQ, the Hospitalized dataset is not only made up of photos but also (and mainly) videos. Knowing that *DeepChest* only takes a selection of images as input, we have to think about a strategy to select the frames of the previously sampled videos (at 15 FPS).

1. Their presence suggests that the clinician saw something suspicious and looked further.

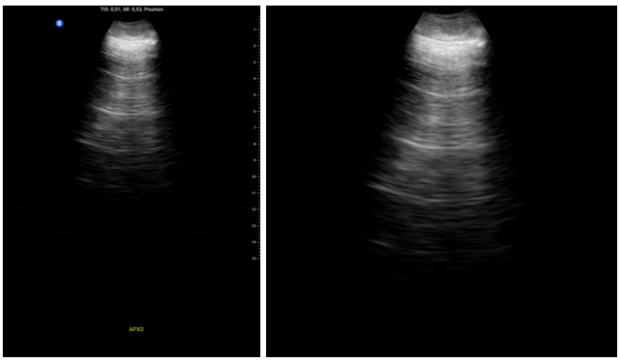


Figure 2.3 – Raw image from ButterflyIQ (Left). Preprocessed image (Right). For this visualization the height ratio between the two images is not respected.

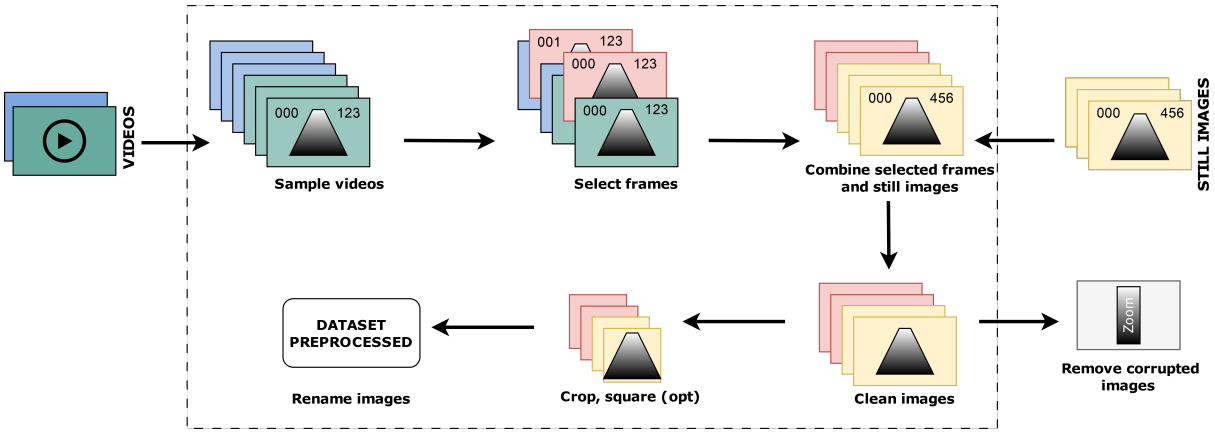


Figure 2.4 – Preprocessing pipeline for the images and videos collected from the Sonosite imaging systems. Cropping and squaring the images is optional.

It is also important to bear in mind that there is a high chance that a (still) image, presumably taken by a clinician, contains information that a random frame taken in a video does not. Some frames in a video may indeed contain more information of diagnostic interest than others.

Several frame selection strategies can be considered:

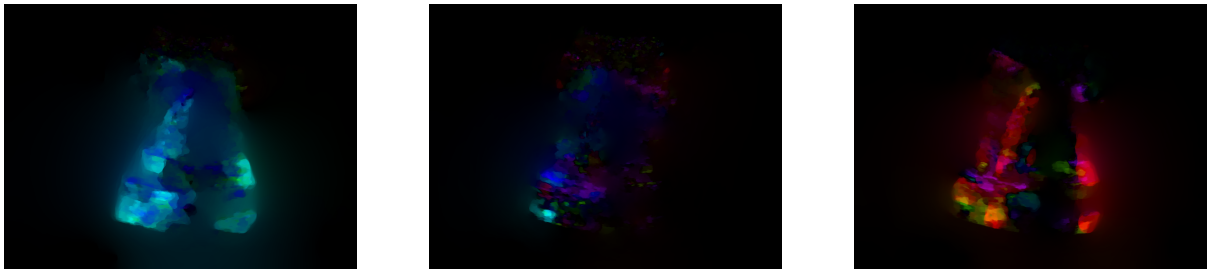
- Select a specific frame in each video, e.g., the 1st one - *probably not optimal as it is difficult to choose a fixed specific frame for each video as they are not the same length. The first frame may be when the operator is not "ready", distracted by the motion of pressing the record button*
- Select a random frame in each video - *simple to implement, avoids building heavy datasets, and is a good way to have a reference.*
- Select several random frames in each video - *using several frames allows a better representation*
- Select several 'synchronous' frames across videos - *perhaps it is interesting to go further and try to align the selected frames between the videos.*

The last point, less straightforward than the previous ones, is explained in more detail in the following sub-section.

Selecting frames using optical flows

On an ultrasound video, horizontal movements from right to left can be seen and we would like to select "synchronous" sequences across videos, i.e, select a number of consecutive frames, say three, which indicate the same movement in each video, say to the left. This corresponds systematically to inspiration or expiration as the probe is held in the same vertical position in each video.

For this purpose, a method based on optical flows [2] has been implemented. The optical flow for each frame was calculated by the TV-L1 algorithm [29, 30]. Optical flows are two-dimensional vector fields (as they represent the horizontal and vertical displacements) and can be represented as an RGB image by a color scale that indicates the direction of movement (see color wheel in the Appendix (Figure A.2). Figure 2.5 shows computed optical flows on three frames extracted from a LUS video.



(a) Movement to the left (b) Stationary (c) Movement to the right

Figure 2.5 – Optical flows computed with TV-L1 algorithm on three frames extracted from a LUS video. Optical flow for a given frame is computed using the previous frame and shows how the pixels have moved between the two frames.

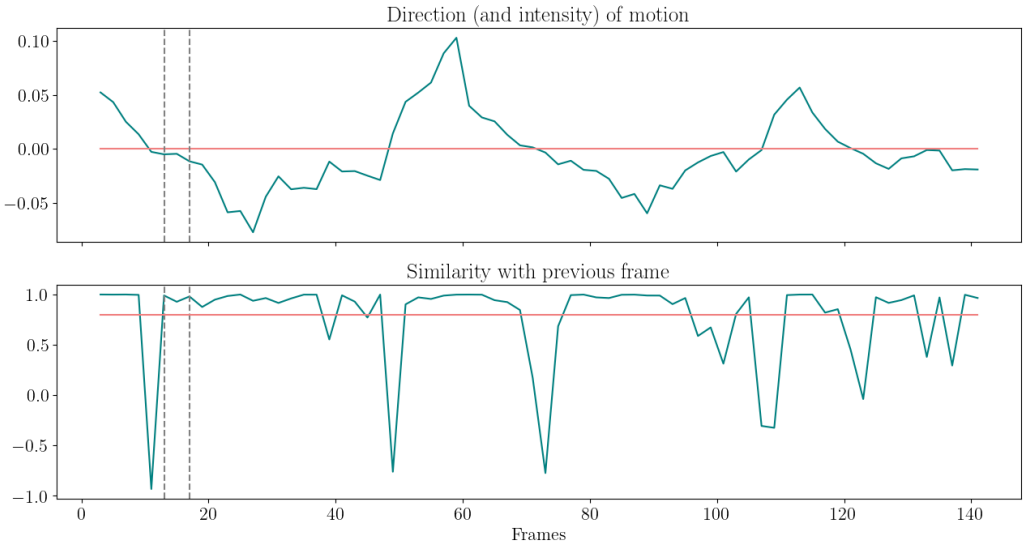


Figure 2.6 – Example of plots obtained with the motion tracking algorithm on a LUS video. For each frame of a video, the algorithm calculates the direction and intensity of the average motion as well as the similarity of that average motion with that of the previous frame.

2. Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of object or camera.

First, we built an algorithm to track the motion in the videos. For each frame i) we determine the direction and intensity of the average motion - left (<0) or right (>0), weak (close to 0) or strong (further from 0) and ii) we check if the average motion is similar (>0.8) or not (<0.8) to the previous frame. Note that the similarity thresholds have been defined somewhat arbitrarily. With the algorithm, for each video we can obtain two plots like those shown in Figure 2.6. The details of the algorithm are given in the Appendix A.3.

We then select for each video the first three consecutive frames that i) have a leftward motion (direction <0) and ii) have a high similarity with the previous frame (similarity > 0.8). In the example in Figure 2.6, the three frames chosen according to this procedure are indicated between the dotted lines. If these conditions cannot be met for a video, the first frame that has a leftward motion (direction <0) is chosen, and if this condition cannot be met either then a random frame is chosen.

Cleaning and cropping images

Images from Sonosite imaging systems are not as easy to clean as images from the ButterflyIQ system. The position of the different elements to be cleaned varies between images, and the nature of these elements - which include text, watermarks and small illustrations - makes the cleaning task challenging. We have developed a relatively efficient, but still imperfect, algorithm that detects the bounding boxes in ultrasound images (roughly cone-shaped) and cleans them by setting all pixels outside this box to 0 (black pixel). An example of the cleaning procedure in shown Figure 2.7. This algorithm can be adapted to crop (and square) a cleaned image as shown in the example in Figure 2.8. Further details about the cleaning/cropping procedures and the bounding box detection algorithm are presented in the Appendix A.4.

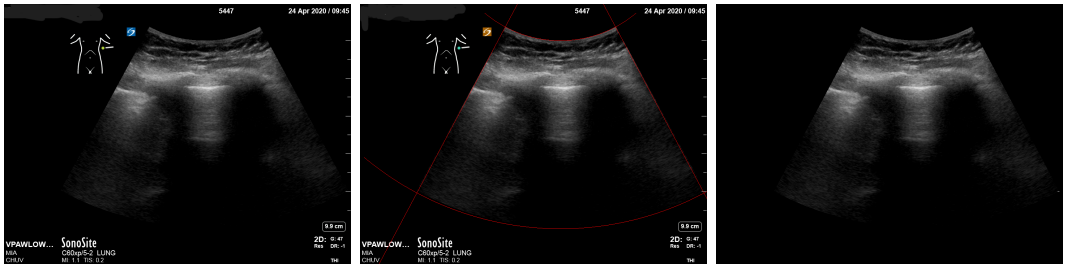


Figure 2.7 – Raw image from Sonosite X-porte (Left). Detected bounding box (Middle). Cleaned image (Right). To ensure anonymity, in this visualization, the patient's name in the upper right-hand corner has been manually masked.

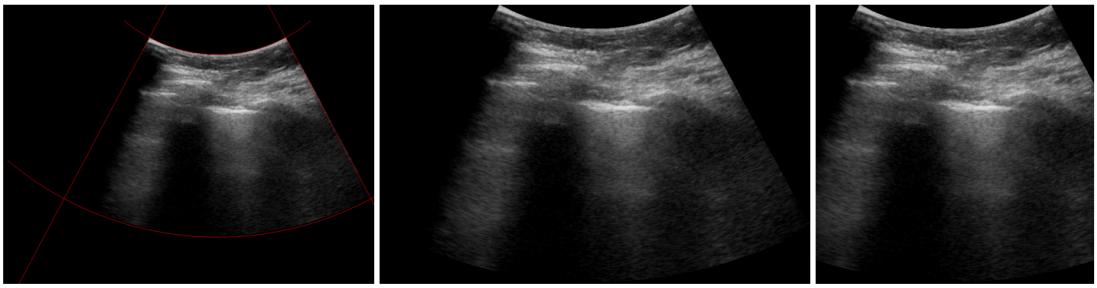


Figure 2.8 – Detected bounding box on cleaned image (Left). Cropped image (Middle). Cropped and squared image (Right). Note that for this visualization the height ratio between the images is not respected.

Removing corrupted images

As with the ButterflyIQ imaging system, we (manually) remove all LUS images that are zoomed in because of the bias that they might introduce in the dataset. Images that have been poorly cleaned (because the bounding box detection algorithm is imperfect) are also removed from the dataset.

2.3 *DeepChest*

2.3.1 Model architecture

DeepChest architecture was developed at iGH. It was designed originally to predict COVID-19 diagnosis from LUS images (collected from ButterflyIQ imaging system). *DeepChest* takes as input the set of LUS images of a patient [3] and their corresponding anatomical position and returns the diagnosis prediction for this patient. The prediction is made at the patient-level as it is clinically more interesting than prediction at the image-level. An overview of *DeepChest* architecture is presented in Figure 2.9. Note that adding the sites embedding is optional and *DeepChest* can be easily adapted to only use the LUS image as input. Details of the three blue blocks in the architecture - namely "Resnet-18", "Embedding layer" and "Attention pooling" - are given in the following subsections.

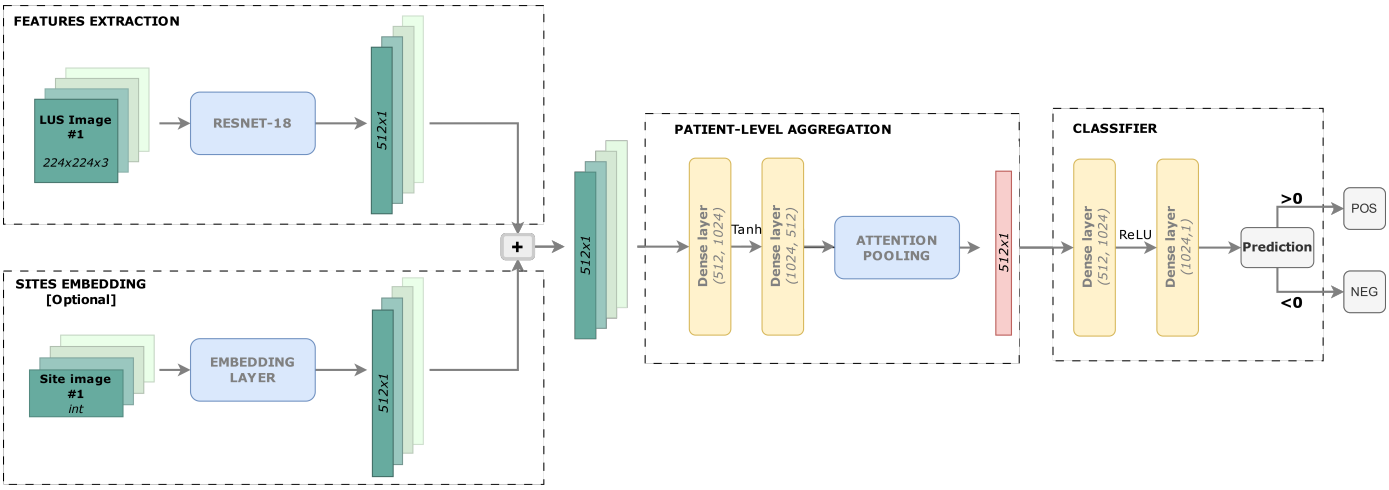


Figure 2.9 – Architecture of *DeepChest*. Adding the sites embedding to the representations obtained at the output of the ResNet is optional.

More formally *DeepChest* (with sites embedding) can be considered as a function DC parameterized by θ which takes as input a set of k images - denoted by X - and their respective anatomical position - s - and returns a scalar x.

$$x = DC_\theta(X, s)$$

The diagnosis is positive (resp. negative) if $x > 0$ (resp. $x < 0$) and the associated probability is given by $\sigma(x)$ where σ is the (standard) logistic function.

3. The number of images might vary between patients.

RESNET-18

ResNet (short for "Residual Network") models were introduced in the paper "Deep Residual Learning for Image Recognition" [31]. This type of deep neural network uses residual connections - also known as "shortcut connections" (Figure 2.10) - to avoid the problem of vanishing gradients in very deep convolutional neural networks (CNNs). This allows ResNets to be much deeper than traditional CNNs and to achieve state-of-the-art performance in visual recognition tasks. *DeepChest* uses a 18-layers residual model (ResNet-18) that was pretrained on ImageNet.

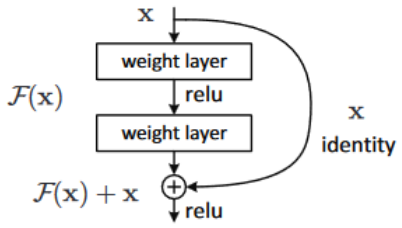


Figure 2.10 – A shortcut connection. Illustration from [31].

Embedding layer

An embedding layer is a type of layer in a neural network that learns to map categorical variables, such as words or identifiers, to continuous vectors of fixed size (embeddings). These embeddings capture the semantic meaning of the variables in a way that is useful for downstream tasks, such as classification or regression. The embedding layer is usually trained by backpropagation, and the resulting embeddings can be used as input features for subsequent layers of the network. In *DeepChest*, the categorical variables are the different possible anatomical sites that are represented by a unique integer identifier and they are embedded in vectors of dimension 512×1.

Attention pooling

Attention pooling is a technique for aggregating a set of input vectors by calculating their weighted average . The weights are learned dynamically based on the relevance of the vector to the task at hand. For *DeepChest*, the set of the images representations (of size 512×1 each) of a given patient obtained at the output of the ResNet - representations to which we may have optionally added the embeddings of the anatomical sites - are aggregated into a single representation (of size 512×1). A graphical representation of the attention pooling in *DeepChest* is presented in the Figure 2.11.

2.3.2 Cross-validation and metrics

Since images from the same patient can be very similar, train/validation/test splits are performed at the patient level and not at the image level as this would leak information. The datasets are split into a train set and test set and 5-folds cross-validation (CV) is performed on the train set. Cross-validation allows to select the hyperparameters but also to obtain 5 models, trained on 5 different train/validation splits. The performance of each configuration is then evaluated on each of the 5 resulting models, and the mean and standard deviation can be calculated.

From a clinical perspective, two evaluation metrics are of particular interest. The first one is the Balanced Accuracy (arithmetic mean between the True Positive Rate - also known as sensitivity - and

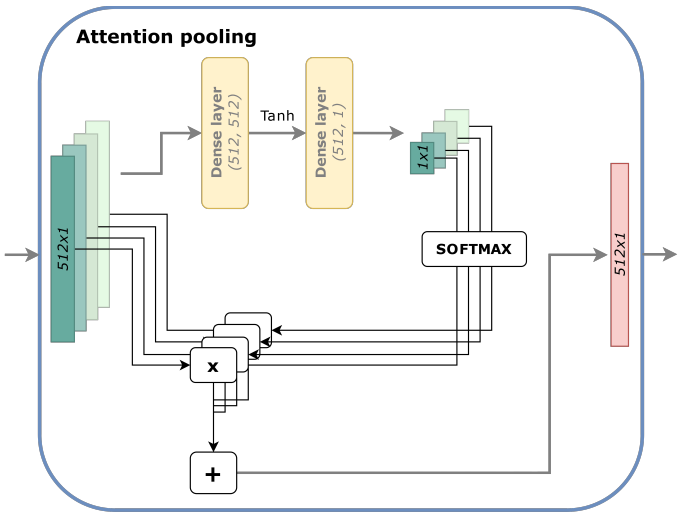


Figure 2.11 – Architecture of the attention pooling module in *DeepChest*

the True Negative Rate - or specificity. The second metric of interest is the Area Under the Curve of the Receiving Operator Characteristic (ROC AUC) which plots the True Positive Rate against the False Positive Rate (computed as 1 - specificity). The ROC AUC has a meaningful interpretation for disease classification from healthy subjects [32].

2.3.3 Training parameters

All the input LUS images are resized (224×224) and normalized before being fed into the network. Furthermore, strictly on the training images, a data augmentation work is done using the *transforms* module of *torchivision* library. Three transforms are applied - namely, ColorJitter, RandomHorizontalFlip and RandomResizedCrop - in order to limit overfitting and maximize generalization.

DeepChest is trained with RAdam [33] and the Binary Cross Entropy Loss (BCE), using an initial learning rate set to 0.001 combined with a scheduler. Details about training parameters are presented in the Appendix under Table A.5.

2.4 Models from clinical tabular data

In addition to LUS images, clinical tabular data were collected for the patients in the TB cohorts. To build a clinical baseline model (i.e. a model that predicts the diagnosis of TB from the clinical tabular data), we use classical machine learning methods for binary classification from numerical and/or categorical data. Their implementation is greatly facilitated by the ready-to-use functions of *Scikit-Learn* and *Pandas* libraries. The general method can be summarized as follow:

1. A preprocessing pipeline is created to impute the few missing values, scale the numerical data (Standard Scaler) and encode the categorical data (Ordinal Encoding).
2. Features are then selected by Recursive Feature Elimination (RFE) [34] and 5-folds CV.
3. A Logistic Regression model is trained with the selected features and the hyperparameters are selected with CV.
4. The model is evaluated on the test data, unseen during training and features/parameters selection.

2.5 Multimodality

While it is possible to build independent models exploiting LUS images and tabular clinical data respectively, it is clinically relevant to consider a model that would exploit the different input modalities to predict the diagnosis of TB - as a clinical expert would certainly do. Several studies have shown that using multimodal models can increase predictive accuracy by an average of 4.2% compared to unimodal models for health diagnosis/prognosis problems [35].

Early, intermediate, and late fusion. One of the major challenges of multimodal learning is the fusion of different modalities which can take place at several levels: before feature extraction, before the classification or after the training of the model (usually referred as 'early', 'intermediate', or 'late' fusion) [36]. In this project, we explore the late fusion strategy, which is very easy to implement, and compare it to the intermediate fusion strategy. We exclude the early fusion strategy because it is more difficult to implement due to the nature of the data to be fused.

Architecture and training. The late fusion strategy uses two previously and independently trained models (*DeepChest* on LUS images and a Logistic Regression model on clinical data) and averages their respective output (probability prediction) to get the final decision. The intermediate fusion strategy trains the modalities jointly, learning a representation from the clinical data and concatenating it with the representations obtained with *DeepChest* before the classification layer (see Figure 2.12). The intermediate fusion model is trained exactly as *Deepchest* - with 5-folds CV and optimized with RAdam and a learning rate of 0.001 (combined with a scheduler).

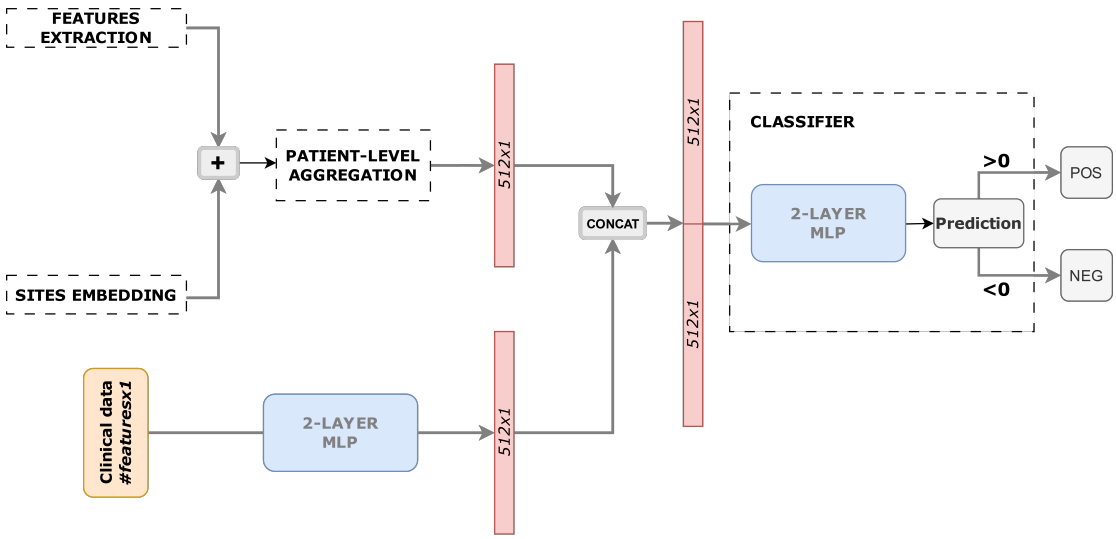


Figure 2.12 – Multimodal architecture with intermediate fusion for TB diagnosis. Learned representations for each modalities are concatenated before the classification head.

3. Experiments

3.1 *DeepChest* for COVID-19 diagnosis

The experiments for COVID-19 diagnosis are summarized in Table 3.2 and more details are given in the following paragraphs.

Table 3.2 – Experiments with *DeepChest* for COVID-19 diagnosis.

Experiment	**#1**	**#2**	**#3**	**#4**
Train on	ER	Hospitalized (1 frame)	Hospitalized (3 frames)	Combined ER/Hospitalized
Test on	ER and Hospitalized	Hospitalized (1 frame)	Hospitalized (3 frames)	ER and Hospitalized
Test size	20% (ER) 25% (Hospitalized)	25%	25%	20% (ER) 25% (Hospitalized)
Pretrained model on ER	Non applicable	- Yes - No	Yes	Yes
Sites embedding	- Yes - No	- Yes - No	No	No
Number of epochs	100	- 100 (Random initialization) - 30 (Pretrained)	30	30
Batch size	32	16	16	16
Positive weight	0.4	1.4	1.4	0.8

ER dataset (#1)

Replication of results. In order to familiarize ourselves with the code and to build our own pretrained models, we reproduce the original results from Schmutz et al. [16] and train *DeepChest* on the ER dataset with and without sites embedding.

Out-of-distribution performance. To assess the generalization performance of *DeepChest*, the trained models from the previous section are evaluated on the Hospitalized test set. Two variations of preprocessing for the Hospitalized dataset images are implemented i) the images are cleaned but not cropped nor squared ii) the images are cleaned, cropped and squared. The idea of the second variation is to try to make the Hospitalized dataset images more consistent with the ER dataset images that were used as training images. This variation of preprocessing is only used in the experiment, in the following images from the Hospitalized dataset are not cropped or squared.

Hospitalized dataset with one frame per video (#2)

We found that *DeepChest* trained on the ER dataset does not seem to generalize well to the Hospitalized dataset. To assess the performance of *DeepChest* when trained on the Hospitalized dataset (with 1 random frame/video), two experiments are set up i) we train *DeepChest* on the Hospitalized dataset from scratch, i.e, the weights are initialized randomly ii) we finetune on the Hospitalized dataset a *DeepChest* model pretrained on the ER dataset. The experiments are carried out with and without sites embedding.

Hospitalized dataset with three frame per video (#3)

In order to further exploit the videos and determine whether using more frames can improve the performance, we build two variations of the Hospitalized dataset using two different frame selection strategies i) we select 3 random frames in each video ii) we select 3 consecutive 'synchronous' frames in each video (see section 2.2.2 for more details). We finetune a pretrained *DeepChest* model (without sites embedding) on both variations of the Hospitalized dataset.

Combined dataset (#4)

To know if it would be possible to obtain a shared representation between the two datasets, we finetune a pretrained *DeepChest* model (without sites embedding) on the dataset obtained by combining the ER dataset and the Hospitalized dataset (with 1 random frame/video).

3.2 *DeepChest* for TB diagnosis

DeepChest vs. baseline models

DeepChest and two baseline models are trained to predict the TB diagnosis on the Benin dataset. The idea of this experiment is to compare the predictive capabilities of *DeepChest* with those of a model exploiting clinical data that is relatively easy to collect and those of a model exploiting LUS interpretations of human experts. For a strict comparison the models were evaluated on the same test set which contains 78 patients (20% of the dataset).

DeepChest. Two instances of *DeepChest* are trained on the Benin dataset - one with sites embedding and the other without. The models are trained using 100 epochs, a batch size of 16 and a positive weight of 1.8.

Clinical baseline model. The clinical model is obtained by training a Logistic Regression model with the clinical data - i.e using demographics data, signs and the result of a rapid HIV test. Features are further selected with RFE (CV) and a regularization strength hyperparameter is tuned with CV.

LUS expert human. For the LUS expert model, the approach is similar - a Logistic Regression model is trained on a selection of features - in this case the interpretations (categorical) of LUS images by an expert clinician (see a description of these interpretations in the Appendix under A.2).

Acquisition optimization for *DeepChest* (with sites embedding)

Image acquisition is expensive as it requires the collection of 14 anatomical views per patient. We would like to know if we can remove one or more anatomical views without it affecting the performance of *DeepChest* too much at inference, i.e. *DeepChest* is trained with images collected from all anatomical sites, but at inference, test set predictions are computed on images collected from different smaller sets of anatomical sites . Note that in this experiment only the 68 patients out of the total 78 in the test set who have at least one image per site are considered for evaluation. In addition, for this experiment, for clinical relevance, the number of sites was reduced from 14 to 6 sites by combining the right and left counterparts of a single site, as well as combining the 4 apical sites into 1 site.

To get an idea of a selection of anatomical sites that would maintain a good model performance we adapted two methods: i) the forward feature selection and ii) the backward feature elimination. Both our adaptations of the forward and backward method exploit the CV splits, i.e., sites are selected by

CV and the performance on the test set is then evaluated. At each iteration, each trained *DeepChest* model corresponding to a train/val split is used to make predictions on its associated validation set (using only the images collected from the set of sites under consideration) and the five resulting prediction vectors are concatenated into a single prediction vector. The forward method starts with an empty set of site and successively adds one site to the set so as to maximize the ROC AUC over the prediction vector while the second method starts with all sites and successively removes one site of the set so as to maximize the ROC AUC over the prediction vector.

Multimodal model for TB diagnosis

It is interesting to know whether using both the LUS images and the clinical data results in a better performing model. Two multimodal architectures derived from *DeepChest* are investigated (see the section 2.5 for more details) - implementing two different fusion strategies.

Late fusion. The trained models are *DeepChest* (with sites embedding) for LUS images and the clinical model for the tabular data.

Intermediate fusion. The multimodal model derived from *DeepChest* takes as input the same tabular data as the clinical model as well as the LUS images. It is trained as *DeepChest* but with 25 epochs only (as training with more epochs did not yield better results).

Out-of-distribution performance of *DeepChest*

To assess the generalization performance of *DeepChest*, we evaluate its performance on an external validation dataset. This dataset consists of images collected from patients attending an outpatient facility in rural South Africa. The patients in this cohort are expected to be more severely ill than those in the main cohort, notably because the prevalence of HIV is higher there than in the urban region of Benin where the patients in the main cohort come from. *DeepChest* as well as the clinical model are evaluated on the patients from the South Africa. However, only a fraction of the patients (73/150) have their Body Mass Index (BMI) available in the tabular data. As this feature is important for the diagnosis of tuberculosis, *DeepChest* and the clinical model were evaluated only on these 73 patients.

4. Results

Unless otherwise specified, all results are presented with **mean ± (unbiased) standard deviation** derived from the 5-fold CV.

4.1 *DeepChest* for COVID-19 diagnosis

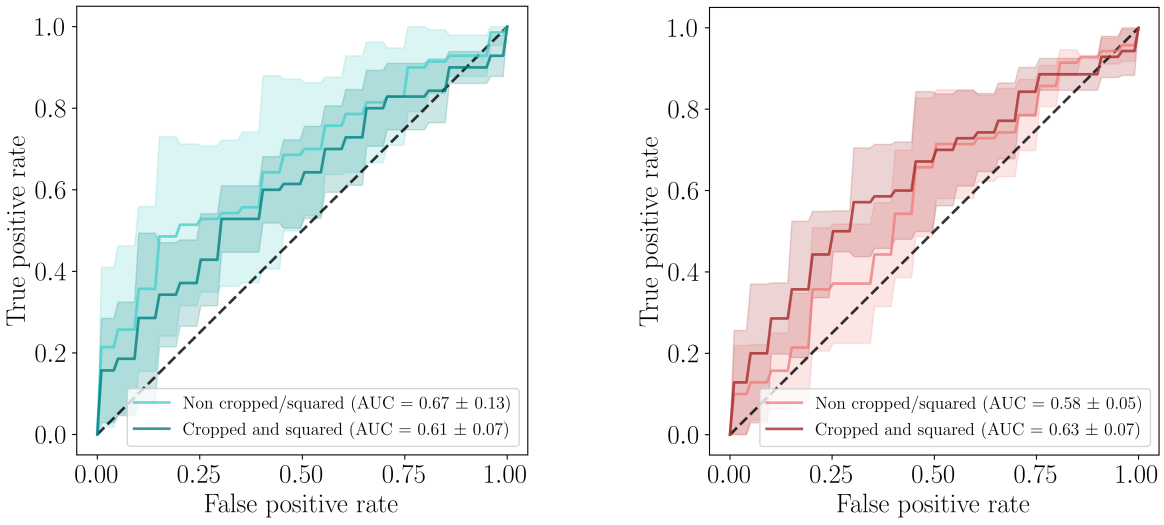
ER dataset

We first tested the generalizability of *DeepChest* - trained on patients recruited in ER - to more severe patient that were hospitalized.

In Table 4.1, we present the results on ER patients. We find that the classification performance of the *DeepChest* is high (with a ROC AUC of 0.87 ± 0.02 for the model without sites embedding) and in agreement with the original results by Schmutz et al. [16]. In Figure 4.1, we present the ROC curves of *DeepChest* (with and without sites embedding) evaluated on the Hospitalized test set. Two variations of image preprocessing (with or without cropping/squaring) are compared. We find that the ROC AUC are low (< 0.70), even with the preprocessing that seeks to conform the images of the Hospitalized dataset to those of the ER dataset. This suggests that *DeepChest* - trained on the ER dataset - does not generalize well to the Hospitalized dataset.

Table 4.1 – Performance of *DeepChest* (trained and tested) on the ER dataset.

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
DeepChest	0.95 ± 0.02	0.66 ± 0.06	0.87 ± 0.02	0.80 ± 0.03
DeepChest with sites embeddings	0.93 ± 0.06	0.70 ± 0.05	0.86 ± 0.01	0.81 ± 0.02



(a) *DeepChest* without sites embedding

(b) *DeepChest* with sites embedding

Figure 4.1 – Performance (ROC curve) of *DeepChest* trained on the ER dataset and tested on the Hospitalized dataset. Two variations of image preprocessing are compared.

Hospitalized dataset with one frame per video

To increase the predictive capabilities of *DeepChest* on hospitalized patients, we fine-tuned a model pretrained on the ER dataset on the Hospitalized dataset and compare results with a *DeepChest* model trained from a random weights initialization.

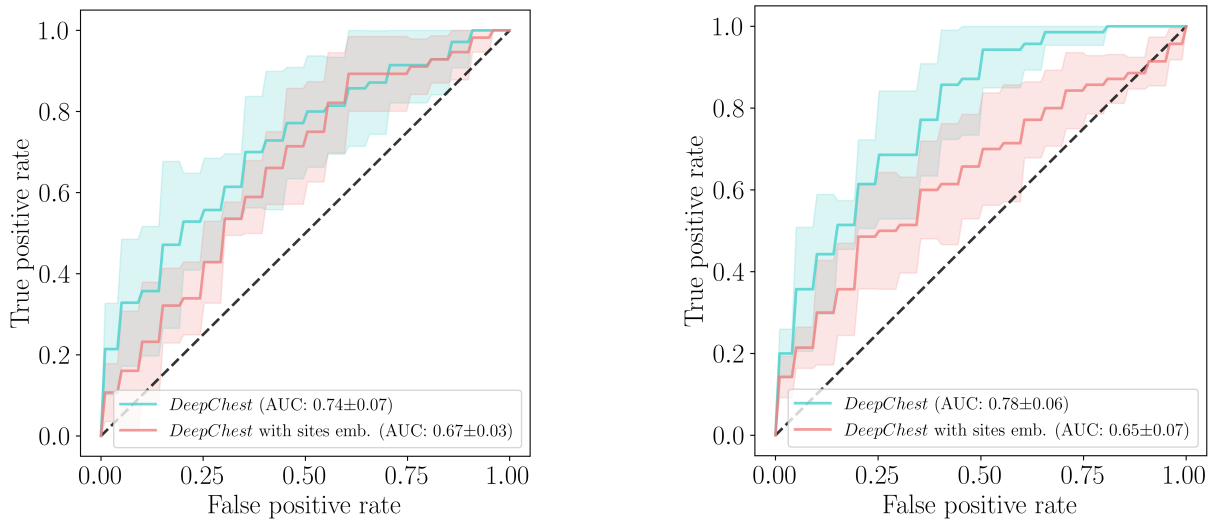
The comparison between training from scratch (Table 4.2) and fine-tuning from a pretrained model (Table 4.2) shows that the best performance is obtained by fine-tuning *DeepChest* without embedding from a pretrained model. This model achieves a ROC AUC of 0.78 ± 0.06. This performance is however below our expectations considering the performance on the ER dataset (Table 4.1). We also note that unlike the results on the ER dataset where the difference with/without sites embedding was not obvious, the present results show a decrease of the ROC AUC when embeddings are added (by 0.13 for the fine-tuned model and by 0.07 for the randomly initialized model).

Table 4.2 – Performance of *DeepChest* trained from scratch on the Hospitalized dataset

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
Deepchest	0.80 ± 0.13	0.50 ± 0.18	0.74 ± 0.07	0.65 ± 0.03
Deepchest with sites embeddings	0.59 ± 0.20	0.65 ± 0.21	0.67 ± 0.03	0.62 ± 0.07

Table 4.3 – Performance of *DeepChest* fine-tuned on the Hospitalized dataset

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
Deepchest	0.71 ± 0.19	0.67 ± 0.15	0.78 ± 0.06	0.69 ± 0.08
Deepchest with sites embeddings	0.44 ± 0.14	0.77 ± 0.18	0.65 ± 0.07	0.61 ± 0.08



(a) *DeepChest* trained from a random weights initialization.

(b) *DeepChest* trained from a model pre-trained on the ER dataset.

Figure 4.2 – Performance (ROC curve) of *DeepChest* on the Hospitalized dataset.

Hospitalized with three frames per video

Unlike the ER dataset which contains only still images selected by a expert clinician, the images in the Hospitalized dataset are randomly extracted from videos, which may explain the poorer performance of *DeepChest* on this dataset. Two variations of the Hospitalized dataset were constructed using two different frame selection strategies to study whether using more frames in each video dataset would improve performance. The results of the experiment, presented in Table 4.4, show an improvement in performance when three random frames are selected in each video (ROC AUC 0.84 ± 0.03) instead of just one (ROC AUC 0.78± 0.06). However our attempt to align frames (synchronous frames) does not improve performance (ROC AUC 0.76 ± 0.04), although it appears to slightly decrease the variance between folds compared to when using one random frame.

Table 4.4 – Performance of *DeepChest* (without sites embedding and pretrained on the ER dataset) on the Hospitalized dataset using different frames selection strategies.

Model - Frame strategy	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
Deepchest - 1 random frame (baseline)	0.71 ± 0.19	0.67 ± 0.15	0.78 ± 0.06	0.69 ± 0.08
Deepchest - 3 random frames	0.79 ± 0.15	0.71 ± 0.14	0.84 ± 0.03	0.75 ± 0.06
Deepchest - 3 synchronous frames	0.74 ± 0.13	0.63 ± 0.12	0.76 ± 0.04	0.69 ± 0.03

Performance when combining both datasets

It was shown that *DeepChest* trained on the ER dataset did not generalize well on the Hospitalized dataset and that it was necessary to fine-tune the model on the latter to obtain better performance. However, by doing this, we no longer have a model that can operate on both datasets at the same time. We therefore wanted to study the performance of *DeepChest* trained on a dataset constructed by combining the ER dataset and the Hospitalized dataset together. The results on the test sets of each dataset, presented in Table 4.5, show that the ROC AUC on each dataset decreases by 0.04 when combining these datasets.

Table 4.5 – Performance of *DeepChest* (without sites embedding and pre-trained on the ER dataset) on the combined dataset.

Model - Test set	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
Deepchest - Hospitalized test set	0.41 ± 0.22	0.79 ± 0.16	0.73 ± 0.07	0.60 ± 0.06
Deepchest - ER test set	0.94 ± 0.03	0.66 ± 0.05	0.83 ± 0.01	0.80 ± 0.02

4.2 *DeepChest* for TB diagnosis

DeepChest vs. baseline models

DeepChest, previously used for COVID-19 detection, is trained on the dataset collected in Benin to predict TB. Two baseline models, obtained respectively from the tabular clinical data and the LUS expert interpretations, are also trained on this dataset. The results for *DeepChest* are presented in Table 4.6 and those for the baselines are presented in Table 4.7. As opposed to COVID, we find that the performance of *Deepchest* is slightly better with sites embeddings. The ROC AUC of *DeepChest* is much higher than that of the LUS expert model and slightly higher than that of the clinical model. *DeepChest* achieves a better specificity and balanced accuracy than the baseline models. The ROC curves associated to these results are presented in Figure 4.3.

Table 4.6 – Performance of *DeepChest* on the Benin test set.

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
DeepChest	0.79 ± 0.11	0.82 ± 0.09	0.90 ± 0.01	0.80 ± 0.02
DeepChest with sites embedding	0.84 ± 0.07	0.84 ± 0.08	0.92 ± 0.02	0.84 ± 0.03

Table 4.7 – Performance of the baseline models on the Benin test set. The selected features for the baseline models are detailed in Table A.4 in Appendix A.6.

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
LUS human expert baseline	0.81 ± 0.02	0.70 ± 0.02	0.84 ± 0.01	0.76 ± 0.02
Clinical baseline	0.83 ± 0.02	0.73 ± 0.03	0.89 ± 0.01	0.78 ± 0.01

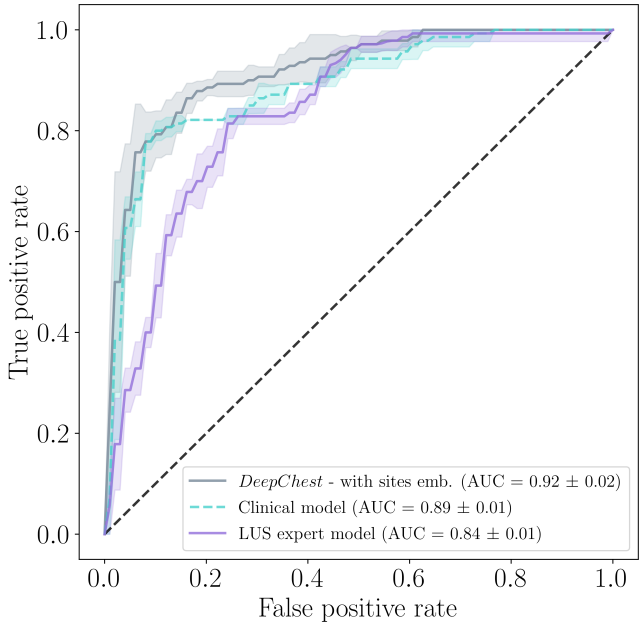
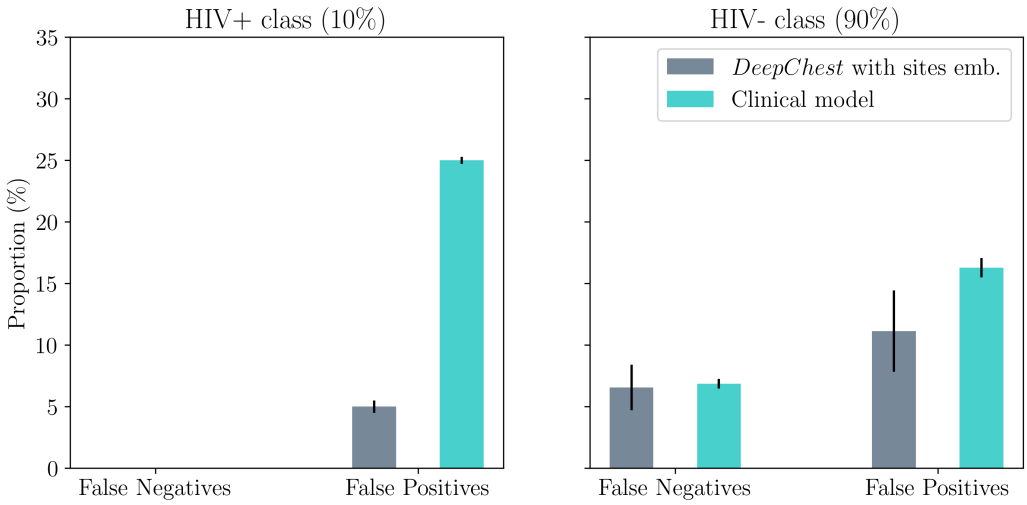


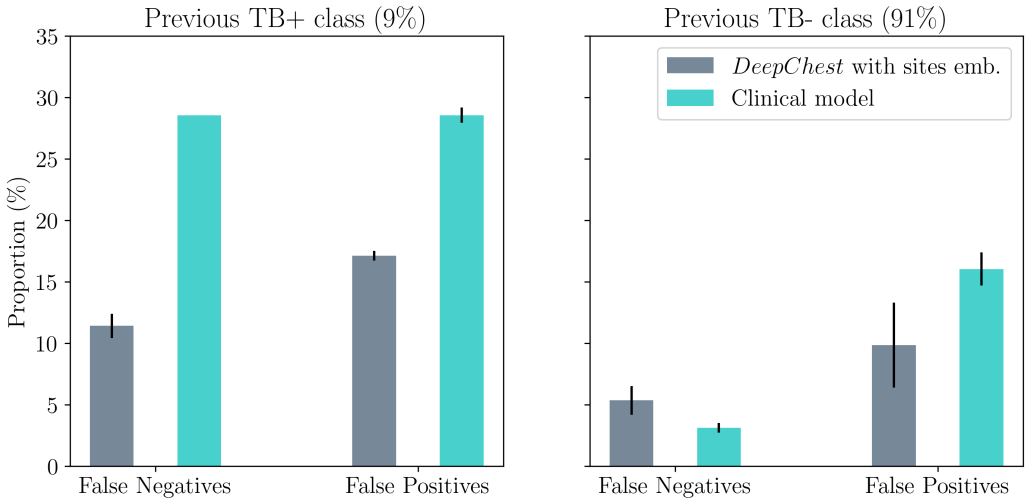
Figure 4.3 – ROC curves of *DeepChest* with sites embedding and LUS human expert and clinical baseline models on the Benin dataset.

Stratified performance

According to Figure 4.3, the classification performance of the clinical baseline model is nearly comparable to that of *DeepChest* with sites embedding. However, a disparity between *DeepChest* and the clinical baseline model is evident when examining performance within specific sub-population classes (stratification). Figure 4.4 reveals that *DeepChest* demonstrates higher accuracy than the clinical baseline model for the HIV-positive population and the population with a history of TB. Clinical models tend to over-diagnose TB in HIV+ patients and both under and over-diagnose TB in patients with prior TB.



(a) HIV

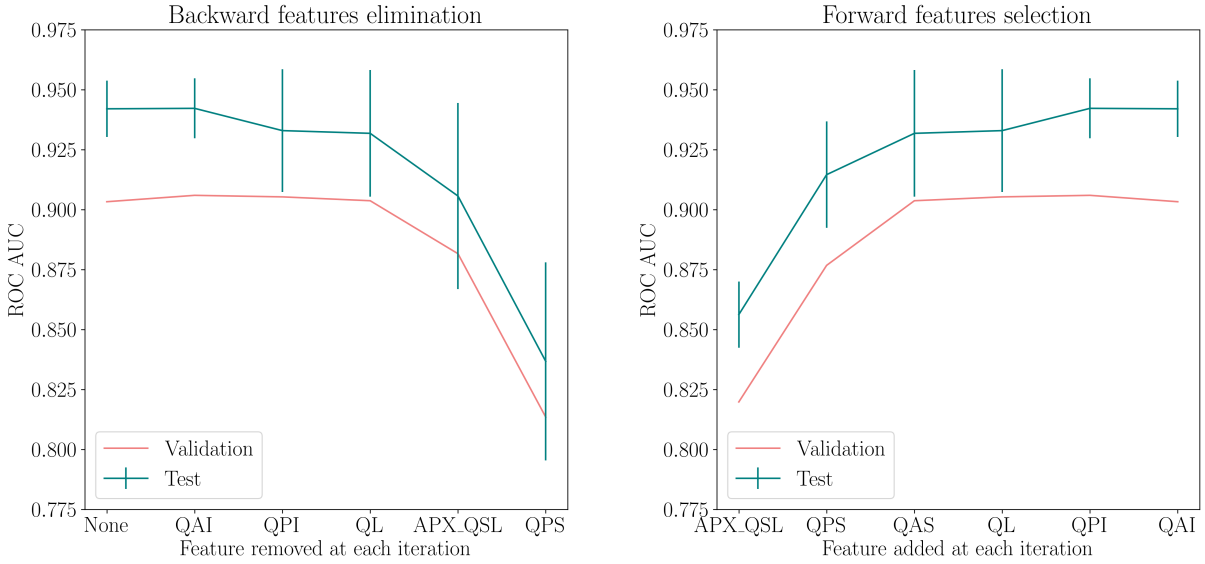


(b) Previous TB

Figure 4.4 – Stratified performance by HIV and Previous TB. The left (resp. right) plot shows the proportion of False Negative and False Positive within the 'HIV positive' (resp. 'HIV negative') class for *DeepChest* with sites embedding and the clinical baseline model. The distribution of the test patients in the two classes is indicated by the percentages written at the top of the plots. The same applies for Previous TB.

Acquisition optimization

The acquisition of the 14 anatomical sites is costly. It was therefore investigated whether sites could be removed at inference without degrading the performance of *DeepChest*. Two methods of site selection - namely forward features selection and backward features elimination - were used. As a reminder, in this experiment, the left and right sites of the lower and upper positions are combined as well as the 4 apical sites, resulting in 6 sites instead of 14. The results are shown in Figure 4.5. Both site selection methods seem to suggest that the APX_QSL (apical), QPS (posterior superior) and QAS (anterior superior) sites are sufficient to maintain the ROC AUC. These sites corresponds to the eight superior sites.



(a) Backward features elimination

(b) Forward features selection

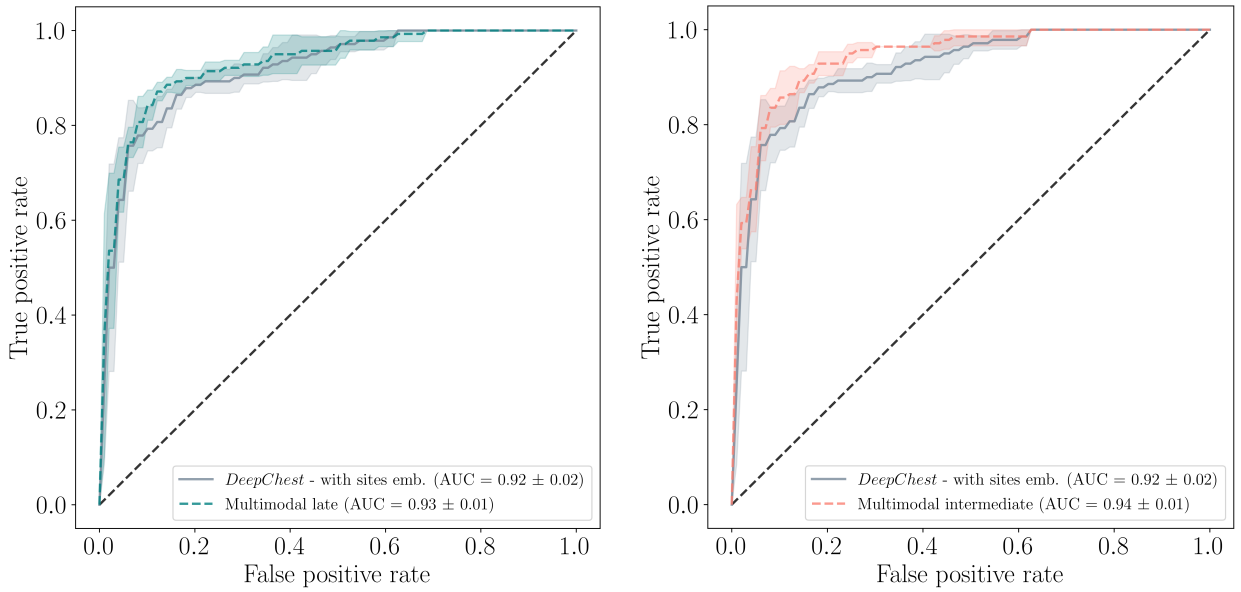
Figure 4.5 – Forward features selection and backward features elimination based on ROC AUC are performed with 5-folds CV. The performance on the test set is evaluated.

Multimodal model

DeepChest was adapted to take tabular clinical data as input in addition to LUS images using two fusion strategies. The performance of resulting multimodal models are presented in Table 4.8. We find that both models performed slightly better than *DeepChest* with an AUC ROC of 0.93 ± 0.01 for the late fusion model and an AUC ROC of 0.94 ± 0.01 for the intermediate fusion model.

Table 4.8 – Performance of multimodal models on the Benin dataset.

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
Late fusion	0.87 ± 0.05	0.86 ± 0.06	0.93 ± 0.01	0.86 ± 0.03
Intermediate fusion	0.91 ± 0.04	0.85 ± 0.06	0.94 ± 0.01	0.88 ± 0.02



(a) Late fusion strategy

(b) Intermediate fusion strategy

Figure 4.6 – ROC curves of the multimodal models.

Out-of-distribution performance

To assess the generalization performance of *DeepChest*, we evaluate its performance on an external validation dataset collected in South Africa. The populations between this dataset and the main dataset are different, particularly in terms of HIV prevalence and history of previous TB (see Table 4.10).

The results presented in Table 4.9 show that *DeepChest* and the clinical model trained on the dataset collected in Benin does not perform well on the dataset collected in South Africa. The performance of *DeepChest* is slightly better than that of a naive model with an ROC AUC of 0.64 ± 0.03 and a balanced accuracy of 0.60 ± 0.06 . The clinical baseline model, on the other hand, does not predict any better than a random model with a ROC AUC of 0.53 ± 0.01 and a balanced accuracy of 0.53 ± 0.04.

Table 4.9 – Performance of *DeepChest* and clinical baseline model trained on dataset collected in Benin and tested on the dataset collected in South Africa.

Model	Sensitivity	Specificity	ROC AUC	Balanced Accuracy
DeepChest with sites embedding	0.41 ± 0.12	0.80 ± 0.07	0.64 ± 0.03	0.60 ± 0.06
Clinical baseline	0.50 ± 0.06	0.56 ± 0.05	0.53 ± 0.01	0.53 ± 0.04

Table 4.10 – Proportion of HIV positive and patients with a history of TB in the population from the Benin and South Africa datasets.

Proportion	Benin	South Africa
HIV+ patients	17%	34%
Previous TB+ patients	14%	31%

5. Discussion

Results and limitations

COVID-19 diagnosis

Generalization to the Hospitalized dataset

Based on the findings, it was determined that *DeepChest* (trained on the ER dataset) did not generalize well to the Hospitalized dataset. However, the reasons for this lack of generalization are difficult to ascertain. It is reasonable to think that the patterns of pathology are different in the two populations (early vs late/severe disease). However, it is likely that the difference in image acquisition could account for the disparity. The images of the Hospitalized dataset are not collected with the same US imaging systems and are quite different from the images of the ER dataset in term of size, bounding box shape, brightness etc. Although we tried to make the images in the Hospitalized dataset more consistent with those in the ER dataset by cropping/squaring, the difference between the two datasets was still noticeable.

Training on the Hospitalized dataset

The study on the Hospitalized dataset is limited by the small size and quality of the dataset. The lack of detailed information on the data acquisition process, particularly the absence of patient metadata, limits the experiments. Additionally, the small size of the dataset, coupled with significant variation in the number of images/videos per patient (ranging from less than 5 to up to 15 videos), contributes to substantial result variance across experiments, which in turn reduces confidence in the results.

When training on the Hospitalized dataset, the performance was lower than the one we observed on the ER dataset. The difference in the selection of input images between the two datasets could explain this observation. In the Hospitalized dataset, the images are extracted randomly from the LUS videos and therefore may not show interesting patterns while the images in the ER dataset have all been selected by an LUS expert. This theory gains support from the observation that performance improves when using three random frames per video instead of a single frame as it is more likely to represent an expert capture. However, it remains challenging to determine why selecting three synchronous frames did not improve performance. It is possible that opting for three consecutive frames with leftward motion might not have been the most effective choice, and selecting three consecutive stationary frames could have been more beneficial. Furthermore, the algorithm itself may be accountable for these results, given its imperfections, such as arbitrary similarity and motion intensity thresholds, as well as its selection criteria that do not consider factors like image brightness level.

TB diagnosis

DeepChest

The performance of *DeepChest* for the diagnosis of TB in Benin are very promising with a ROC AUC superior to 90%. We found that the performance is a bit higher than the performance of *DeepChest* for the diagnosis of COVID-19, notably regarding the specificity. The results also suggest that using

sites embedding in *DeepChest* yields to slightly better results, in contrast to what was observed in the COVID-19 datasets. However, given the margin of error, further experiments are required to confirm these results.

Clinical and LUS expert baseline models

The findings suggest that *DeepChest* performs better than the model built from LUS image interpretations by human experts. For the clinical model this conclusion is less obvious, *DeepChest* has a superior balanced accuracy but the ROC AUC of the two models are very similar. However, *DeepChest* has the significant advantage of being cheap and easy to use. This is not quite the case with the clinical model, which uses the result of an HIV test - data which is not particularly 'cheap' and requires equipment. In addition, the stratified performance shows that the clinical model is less fair to the HIV-positive population and the population with a history of TB than *DeepChest*, which gives the latter an advantage. The performance of the clinical model exceeds our expectations but the results should be treated with caution. We observed, when changing the random seed (and therefore the train/test split), some variability in the performance of the clinical model (which can drastically decrease), reducing confidence in the results presented.

Multimodal model

The findings on the multimodal model are encouraging, but need to be considered in the light of what has just been explained. The clinical model performs 'too well' compared with what is expected, perhaps because of a lucky train/test split.

Acquisition optimization

The results of the procedure aimed at optimizing the acquisition of anatomical sites showed that the eight superior sites may be sufficient to maintain a high performance, and therefore by extension that tuberculosis markers would tend to be located in the upper lung regions. These results are in agreement with those of the LUS expert model. Indeed, the anatomical sites selected by RFE for this model (presented in Table A.4 in the Appendix) are the same as those selected in the acquisition optimization procedure.

Out-of-distribution performance

The performance of *DeepChest* on the external validation dataset is quite low as we expected. As the prevalence of HIV is much higher in South Africa, patients suffering from tuberculosis (and other diseases) are often much more severely ill due to immunocompromise. The disease patterns learned by *DeepChest* from the Benin dataset are therefore not similar to those of patients in South Africa. This underlines the strong dependence of *DeepChest* on the population it is trained on.

Future work

On the Hospitalized dataset, it is difficult to consider future work because the dataset is not large enough and does not contain any patient metadata, which is essential for a complete study. However, this study has highlighted the importance of developing a frame selection method for the LUS video

or a model that takes a LUS video directly as input, thereby dispensing with the expertise required to select still images. This could be useful for other studies using LUS video and in particular for the TB study, as still images are manually selected from the LUS video, making it possible to consider using the data to train a deep learning model that would learn to select the best frames from a video.

DeepChest performs reasonably well on the Benin dataset, but there's still plenty of room for improvement, either by tweaking the model slightly or by further optimizing the hyperparameters.

The generalization of *DeepChest* to the South Africa dataset is weak, underlining the need to work on the generalization of *DeepChest*. By increasing the size of the South Africa dataset, it would be interesting to train *DeepChest* on it and test its generalization to the Benin dataset. In addition, we should consider training a model that could perform well on both datasets. Working on cross-device generalization is also an interesting research prospect.

The multimodal model built in this project is intended to serve as a reference. More complex multimodal models, seeking to combine the different modalities as effectively as possible, are being studied in the laboratory and may soon be used as part of the TB study.

A final and very important point is the interpretability of *DeepChest*. A model whose decisions are not well understood cannot be used in a clinical setting, given the stakes involved. Popular interpretability methods such as Grad-CAM or LIME have major limitations and further research is needed to find ways of making *DeepChest* more interpretable.

Conclusion

This study highlighted the potential of using deep learning to automate the interpretation of lung ultrasound images for the detection of COVID-19 and tuberculosis and maybe other lung diseases. The COVID-19 study paves the way for studies proposing models that use LUS video directly, deployable in regions where LUS expertise is scarce. On the other hand, the TB study has demonstrated the highly promising performance of a deep learning model in diagnosing TB based on LUS images. This solution is not only cost-effective but also user-friendly, aligning with the recommendations of the World Health Organization that advocate for research on new and accessible detection tools as part of its strategy to eradicate TB in the coming years. However, the hesitant performance observed on an external validation dataset as well as the lack of interpretability serve as a reminder that there is still ample room for improvement in this field.

References

[1] European center for disease prevention and control. *COVID-19 testing strategies and objectives*. ECDC: Stockholm; 2020, 15 September 2020.

[2] Zeeshan Sidiq, M. Hanif, Kaushal Kumar Dwivedi, and K.K. Chopra. Benefits and limitations of serological assays in covid-19 infection. *Indian Journal of Tuberculosis*, 67(4, Supplement):S163–S166, 2020. Special Issue on Tuberculosis and COVID-19.

[3] L.J. Krüger, M. Gaeddert, L. Köppel, L. E. Brümmer, C. Gottschalk, I.B. Miranda, P. Schnitzler, H.G. Kräusslich, A.K. Lindner, O. Nikolai, F.P. Mockenhaupt, J. Seybold, V.M. Corman, C. Drosten, N.R. Pollock, A.I. Cubas-Atienzar, K. Kontogianni, A. Collins, A. H. Wright, B. Knorr, A. Welker, M. de Vos, J.A. Sacks, E.R. Adams, C.M. Denkinger, and . Evaluation of the accuracy, ease of use and limit of detection of novel, rapid, antigen-detecting point-of-care diagnostics for sars-cov-2. *medRxiv*, 2020.

[4] World Health Organization. *Global tuberculosis report 2021*. 2021.

[5] World Health Organization. *The End TB Strategy*. 2015.

[6] World Health Organization. *WHO consolidated guidelines on tuberculosis: module 3: diagnosis: rapid diagnostics for tuberculosis detection, 2021 update*. 2021.

[7] World Health Organization. *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. 2021.

[8] Ali Ebrahimi, Mahmoud Yousefifard, Hossein Kazemi, Hamid Reza Rasouli, Hadi Asady, Ali Jafari, and Mostafa Hosseini. Diagnostic accuracy of chest ultrasonography versus chest radiography for identification of pneumothorax: A systematic review and meta-analysis. *Tanaffos*, 13:29–40, 09 2014.

[9] Daniel A Lichtenstein, Ivan Goldstein, Eric Mourgeon, Philippe Cluzel, Philippe A. Grenier, and J. J. Rouby. Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome. *Anesthesiology*, 100:9–15, 2004.

[10] Anna Maw, Ahmed Hassanin, P. Ho, Matthew McInnes, Angela Moss, Elizabeth Juarez, Nilam Soni, Marcelo Miglioranza, Elke Platz, Kristen DeSanto, Anthony Sertich, Gerald Salame, and Stacie Daugherty. Diagnostic accuracy of point-of-care lung ultrasonography and chest radiography in adults with symptoms suggestive of acute decompensated heart failure: A systematic review and meta-analysis. *JAMA Network Open*, 2:e190703, 03 2019.

[11] Yogendra Amatya, Jordan Rupp, Frances Russell, Jason Saunders, Brian Bales, and Darlene House. Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting. *International Journal of Emergency Medicine*, 11, 03 2018.

[12] Véronique Suttels, Jacques Daniel Du Toit, Arnauld Attannon Fiogbé, Ablo Prudence Wachinou, Brice Guendehou, Frédéric Alovokpinhou, Péricles Toukoui, Aboudou Rassisou Hada, Fadyl Sefou, Prudence Vinasse, Ginette Makpemikpa, Diane Capo-chichi, Elena Garcia, Thomas Brahier, Kristina Keitel, Khadidia Ouattara, Yacouba Cissoko, Seydina Alioune Beye, Pierre-André Mans, Gildas Agodokpessi, Noémie Boillat-Blanco, and Mary Anne Hartley. Point-of-care ultrasound for tuberculosis management in sub-saharan africa—a balanced swot analysis. *International Journal of Infectious Diseases*, 123:46–51, 2022.

[13] Siméon Schaad, Thomas Brahier, Mary-Anne Hartley, Jean-Baptiste Cordonnier, Luca Bosso, Tanguy Espejo, Olivier Pantet, Olivier Hugli, Pierre-Nicolas Carron, Jean-Yves Meuwly, and

Noemie Boillat-Blanco. Point-of-care lung ultrasonography for early identification of mild covid-19: a prospective cohort of outpatients in a swiss screening center. *medRxiv*, 2021.

[14] Sachita Shah, Blaise Bellows, Adeyinka Adedipe, Jodie Totten, Brandon Backlund, and Dana Sajed. Perceived barriers in the use of ultrasound in developing countries. *Critical ultrasound journal*, 7:28, 12 2015.

[15] Jing Wang, Xiaofeng Yang, Boran Zhou, James J. Sohn, Jun Zhou, Jesse T. Jacob, Kristin A. Higgins, Jeffrey D. Bradley, and Tian Liu. Review of machine learning in lung ultrasound in covid-19 pandemic. *Journal of Imaging*, 8(3):65, Mar 2022.

[16] Hugo Schmutz, Jean-Baptiste Cordonnier, Amir Rezaie Thomas Brahier, Makhmutova, Jean-Yves Meuwly, Olivier Pantet, Marie-Josée Brochu Vez, Olivier Huglia, Martin Jaggi, Noemie Boillat-Blanco, and Mary-Anne Hartley. Automated detection and risk stratification of covid-19 from lung ultrasound : A deep learning model trained on 296 adults with pneumonia at a swiss emergency unit. In progress.

[17] Imran Ul Haq. An overview of deep learning in medical imaging, 2022.

[18] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, aug 2021.

[19] Mustapha Oloko-Oba and Serestina Viriri. A systematic review of deep learning techniques for tuberculosis detection from chest radiograph. *Frontiers in Medicine*, 9, 03 2022.

[20] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4:475–7, 12 2014.

[21] Tawsifur Rahman, Amith Khandakar, Muhammad Kadir, Khandaker Islam, Forhad Khandakar, Rashid Mazhar, Tahir Hamid, Mohammad Islam, Saad Kashem, Mohamed Ayari, and Muhammad Chowdhury. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 10 2020.

[22] Jannis Born, Gabriel Brändle, Manuel Cossio, Marion Disdier, Julie Goulet, Jérémie Roulin, and Nina Wiedemann. Pocovid-net: Automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus), 2021.

[23] Julia Diaz-Escobar, Nelson E. Ordóñez-Guillén, Salvador Villarreal-Reyes, Alejandro Galaviz-Mosqueda, Vitaly Kober, Raúl Rivera-Rodriguez, and Jose E. Lozano Rizk. Deep-learning based detection of covid-19 using lung ultrasound imagery. *PLOS ONE*, 16(8):1–21, 08 2021.

[24] Salehe Ebadi, Deepa Krishnaswamy, Seyed Bolouri, Dornoosh Zonoobi, Russ Greiner, Nathaniel Meuser-Herr, Jacob Jaremko, Jeevesh Kapur, Michelle Noga, and Kumaradevan Punithakumar. Automated detection of pneumonia in lung ultrasound using deep video classification for covid-19. *Informatics in Medicine Unlocked*, 25:100687, 08 2021.

[25] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, and Libertario Demi. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Transactions on Medical Imaging*, PP, 05 2020.

[26] Thomas Brahier, Jean-Yves Meuwly, Olivier Pantet, Marie-Josee Brochu Vez, Helene Gerhard Donnet, Mary-Anne Hartley, Olivier Hugli, and Noemie Boillat-Blanco. Lung ultrasonography for risk stratification in patients with coronavirus disease 2019 (covid-19): A prospective observational cohort study. *Clinical Infectious Diseases*, 73(11):E4189–E4196, 2021.

[27] Véronique Suttels, Prudence Wachinou, Jacques Du Toit, Noémie Boillat-Blanco, and Mary-Anne Hartley. Ultrasound for point-of-care sputum-free tuberculosis detection: Building collaborative standardized image-banks. *EBioMedicine*, 81:104078, July 2022.

[28] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[29] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *Proceedings 29th DAGM Symposium "Pattern Recognition"*, pages 214–223. Springer, 2007. 29th DAGM Symposium on Pattern Recognition : DAGM 2007 ; Conference date: 12-09-2007 Through 14-09-2007.

[30] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3:137–150, 2013. `https://doi.org/10.5201/ipol.2013.26`.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[32] Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4:627–635, 09 2013.

[33] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265, 2019.

[34] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 01 2002.

[35] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health, 2022.

[36] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017.

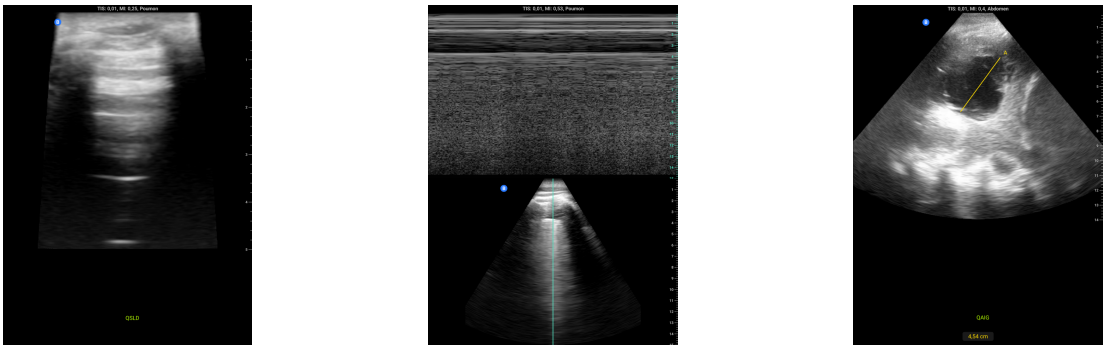
A. Appendix

A.1 Clinical tabular data : Expert human interpretations

Table A.2 – For each LUS image collected from one of the 14 anatomical sites, the LUS human expert gives an interpretation encoded by an integer ranging from 0 (normal) to 6 (very serious). The expert also indicate if there is a pleural effusion (unsual amount of fluid around the lungs) or not.

	Interpretation	Value
Anatomic sites (14)	*Normal Pattern (A-ligns)*	0
	Pattern with $\geq$ 3 B-lines per field	1
	Pattern with coalescing B-lines	2
	Pattern with small consolidations and/or subpleural nodules (< 1cm height)	3
	Consolidation of $\geq$ 1 cm in height	4
	Pattern A' (pneumothorax)	5
	Pleural effusion	6
Pleural effusion	*No*	0
	Yes	1

A.2 Corrupted images from ButterflyIQ



(a) Zoomed image. (b) Image taken in M-mode. (c) Image with ruler.

Figure A.1 – Examples of corrupted images that need to be removed from the datasets as they might introduce a bias.

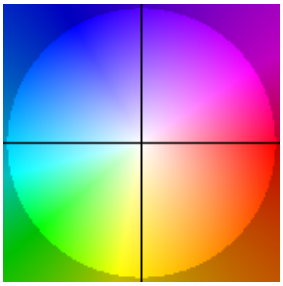


Figure A.2 – Color wheel for optical flows visualization. The color indicates the direction and intensity of the motion

A.3 Selecting 'synchronous' frames using optical flows

The motion tracking algorithm is described below and uses the cosine similarty to measure similarity and the projection on the horizontal axis (x) to measure the direction and intensity of the (mean) motion . For each video, the algorithm returns two lists - **sim** and **proj** - which can be plotted and used to select frames.

Algorithm 1 Tracking the motion in a sampled video

Input: Sampled frames (f_1, f_2, ... , f_N)

1. Initialize two lists of size N: one to store the similarity (**sim**) and the other to store the projection (**proj**). Set **sim**[1] = 0 and **proj**[1] = 0.
2. Set the initial average flow $\overline{\text{prev\_flow}}$ to $[0, 0]$.
3. for $i \in \{2, 3, \ldots, N\}$:
   (a) Compute the horizontal (flow_x,) and vertical (flow_y) displacements between frame f_i and frame f_{i-1} using TV-L1 algorithm.
   (b) Compute the mean over all the pixels of the vertical ($\overline{\text{flow}_x}$) and horizontal ($\overline{\text{flow}_y}$) displacements.
   (c) Compute the cosine similarity c between $[\overline{\text{flow}_x}, \overline{\text{flow}_y}]$ and average previous flow $\overline{\text{prev\_flow}}$.
   (d) Set **sim**$[i] = c$ and **proj**$[i] = \overline{\text{flow}_x}$.
   (e) Update $\overline{\text{prev\_flow}} = [\overline{\text{flow}_x}, \overline{\text{flow}_y}]$.

Output: **sim** and **proj**

A.4 Cleaning/cropping algorithm for Sonosite LUS images

To detect the bounding box of an image:

1. Detect 3 points on the upper arc : right/left ends and lowest point on the arc.
2. Compute the center and radius of the associated circle. Draw upper arc. Compute and draw right (resp. left) line between the center and the right (resp. left) end.
3. Compute the radius of the lower arc (hack). Draw lower arc.

To clean an image:

4. Set all pixel outside the detected bounding box to 0.

To crop an image:

5. Crop along right the horizontal line passing through right and left end of the upper arc.
6. Compute the horizontal tangent of the lower arc. Crop horizontally along this tangent.
7. Compute the intersections between the lower arc and the two lines from step 2. Compute the vertical lines passing throught these intersections. Crop along theses two lines.

Note that the detection thresholds for the 3 points on the upper arc (step 1) are different for Sonosite X-porte and Sonosite M-turbo images.

A.5 Training parameters

Training parameters used to train *DeepChest* are presented in Table A.3. It can be noted that the *DeepChest* model shown in the Figure 2.9 does not contain a sigmoid activation at the output of the network, as one might expect for a binary classification task. This is because *DeepChest* uses the Binary Cross Entropy (BCE) loss directly combined with a Sigmoid layer for reasons of numerical stability. The loss can be formalised as follow:

$$L(x, y) = -\frac{1}{N}\sum_{i=1}^{N} \ell_n, \qquad \ell_n = p_c y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))$$

where N is the batch size, x (resp. y) is a vector of size N containing the outputs of *DeepChest* (resp. the ground-truth labels) for the patients in the batch and p_c ('positive weight') is a weight to trade off recall and precision.

Table A.3 – *DeepChest* training parameters

Loss function	*BCEWithLogitsLoss*[1] (positive weight = $\frac{\#\text{of negatives samples}}{\#\text{of positives samples}}$)				
Optimization algorithm	RAdam [33]				
	Learning rate	β_1	β_2	ε	Weight decay
	0.001	0.9	0.999	$1e-8$	0.0
Scheduler	ExponentialLR ($\gamma = 0.98$)				

[1] `https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`

A.6 Features selected with RFE

Table A.4 – Features selected with RFE (CV) for the baseline models

Cheap clinical	sex, bmi, hiv, previous tb, hypertension, diabetes, heart rate, systolic blood pressure
LUS expert human	APX_QSL (apical), QAS (anterior superior), QPS (posterior superior), pleural effusion

Master's Theses in Mathematical Sciences 2023:E45
ISSN 1404-6342

LUTFMA-3514-2023

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/